CLAIMS

1. A method of executing a query for at least one document similar to a specified document, the method comprising:

receiving the query;

forming a reduced query document based on ranks of terms in the specified document;

generating a modified query based on the query and the reduced query document;

executing the modified query on a data repository to generate a set of results; and

providing a result to a user interface.

2. A method in accordance with claim 1, wherein the result comprises a document that is similar to the specified document.

3. A method in accordance with claim 1, wherein the result comprises a list of references to documents that are similar to the specified document.

4. A method in accordance with claim 1, wherein the result indicates that no similar document was found in the data repository.

5. A method in accordance with claim 1, wherein forming the reduced query document based on ranks of terms in the specified document excludes terms that are less selective.

6. A method in accordance with claim 1, wherein forming the reduced query document based on ranks of terms in the specified document comprises:

calculating a rank of at least one term in the specified query document;

calculating a square of each rank;

calculating a normalized rank for each square;

sorting a list of normalized ranks including the normalized rank;

calculating a partial sum for each normalized rank in the list of normalized ranks; and

including, in the reduced query document, terms corresponding to a partial sum above a threshold value.

7. A method in accordance with claim 1, wherein the data repository is modeled in accordance with a vector space model and executing the modified query comprises calculating the similarity of the reduced query document with a comparison document in the data repository in accordance with the function $Q \cdot D \ / \ |Q| * |D|$ where Q is the reduced query document, D is the comparison document, $Q \cdot D$ is a scalar product of column vectors corresponding to each document such that each column is a vector including ranks of terms in the documents, and $|Q| * |D|$ is a normalization factor.

8. A method in accordance with claim 7, wherein the normalization factor is the product of the norms of the column vectors corresponding to each document calculated in accordance with the equation $\text{sqrt}(q1^2 + \ldots + qT^2) * \text{sqrt}(d1^2 + \ldots + dT^2)$ where sqrt signifies square root, q1 through qT are ranks of terms in the reduced query document, d1 through dT are ranks of terms in the comparison document, and T is the number of terms in an index of document vectors generated in accordance with the vector space model of the data repository.

9. A method in accordance with claim 7, wherein the scalar product of the column vectors is calculated in accordance with the equation $(q1.d1 + q2.d2 + \ldots + qT.dT)$ where q1 through qT are ranks of terms in the reduced query document, d1 through dT are ranks of terms in the comparison document, and T is the number of terms in an index of document vectors generated in accordance with the vector space model of the data repository.

10. An information management system, the system comprising:

a data repository, wherein the data repository is configured to store documents; and

a program for executing queries on the data repository, wherein the program is operative to:

receive a query for at least one document similar to a specified document;

form a reduced query document based on ranks of terms in the specified document;

generate a modified query based on the query and the reduced query document;

execute the modified query on the data repository to generate a set of results; and

provide a result to a user interface.

11. An information management system in accordance with claim 10, wherein the result comprises a document that is similar to the specified document.

12. An information management system in accordance with claim 10, wherein the result comprises a list of references to documents that are similar to the specified document.

13. An information management system in accordance with claim 10, wherein the result indicates that no similar document was found in the data repository.

14. An information management system in accordance with claim 10, wherein the operation of forming the reduced query document based on ranks of terms in the specified document excludes terms that are less selective.

15. An information management system in accordance with claim 10, wherein the operation of forming the reduced query document based on ranks of terms in the specified document comprises:

calculating a rank of at least one term in the specified query document;

calculating a square of each rank;

calculating a normalized rank for each square;

sorting a list of normalized ranks including the normalized rank;

calculating a partial sum for each normalized rank in the list of normalized ranks; and

including, in the reduced query document, terms corresponding to a partial sum above a threshold value.

16. An information management system in accordance with claim 15, wherein the data repository is modeled in accordance with a vector space model and the operation of executing the modified query comprises calculating the similarity of the reduced query document with a comparison document in the data repository in accordance with the function $Q \cdot D \ / \ |Q| * |D|$ where Q is the reduced query document, D is the comparison document, $Q \cdot D$ is a scalar product of column vectors corresponding to each document such that each column is a vector including ranks of terms in the documents, and $|Q| * |D|$ is a normalization factor.

17. An information management system in accordance with claim 15, wherein the normalization factor is the product of the norms of the column vectors corresponding to each document calculated in accordance with the equation sqrt(q1^2 + … + qT^2) * sqrt(d1^2 + ... + dT^2) where sqrt signifies square root, q1 through qT are ranks of terms in the reduced query document, d1 through dT are ranks of terms in the comparison document, and T is the number of terms in an index of document vectors generated in accordance with the vector space model of the data repository.

18. An information management system in accordance with claim 15, wherein the scalar product of the column vectors is calculated in accordance with the equation (q1.d1 + q2.d2 + ... + qT.dT) where q1 through qT are ranks of terms in the reduced query document, d1 through dT are ranks of terms in the comparison document, and T is the number of terms in an index of document vectors generated in accordance with the vector space model of the data repository.